(12) UK Patent Application (19) GB (11) 2 320 112 (13) A

(43) Date of A Publication 10.06.1998

(21) Application No 9625454.5

(22) Date of Filing 07.12.1996

(51) INT CL[6]
G06F 13/14

(52) UK CL (Edition P)
G4A AFN

(56) Documents Cited
US 4257099 A US 4050095 A

(58) Field of Search
UK CL (Edition O) G4A AFGN AFN
INT CL[6] G06F 13/00 13/14
Online: WPI

(71) Applicant(s)
International Business Machines Corporation
(Incorporated in USA - New York)
Armonk, New York 10504, United States of America

(72) Inventor(s)
Adrian Mark Colyer

(74) Agent and/or Address for Service
R D Moss
IBM United Kingdom Limited, Intellectual Property Department, Mail Point 110, Hursley Park, WINCHESTER, Hampshire, SO21 2JN, United Kingdom

(54) High-availability computer server system

(57) A high-availability computer server system (36) capable of serving a large number of requests received from a plurality of computer client devices connected through a network to said server system, said requests specifically identifying said server system, comprises a messaging and queuing unit (31) having an input connected to said network upon which said requests identifying said server system are received, and an output; and a plurality of server units (32) connected in parallel to said output of said messaging and queuing unit. The server may be used as a server on the World Wide Web.

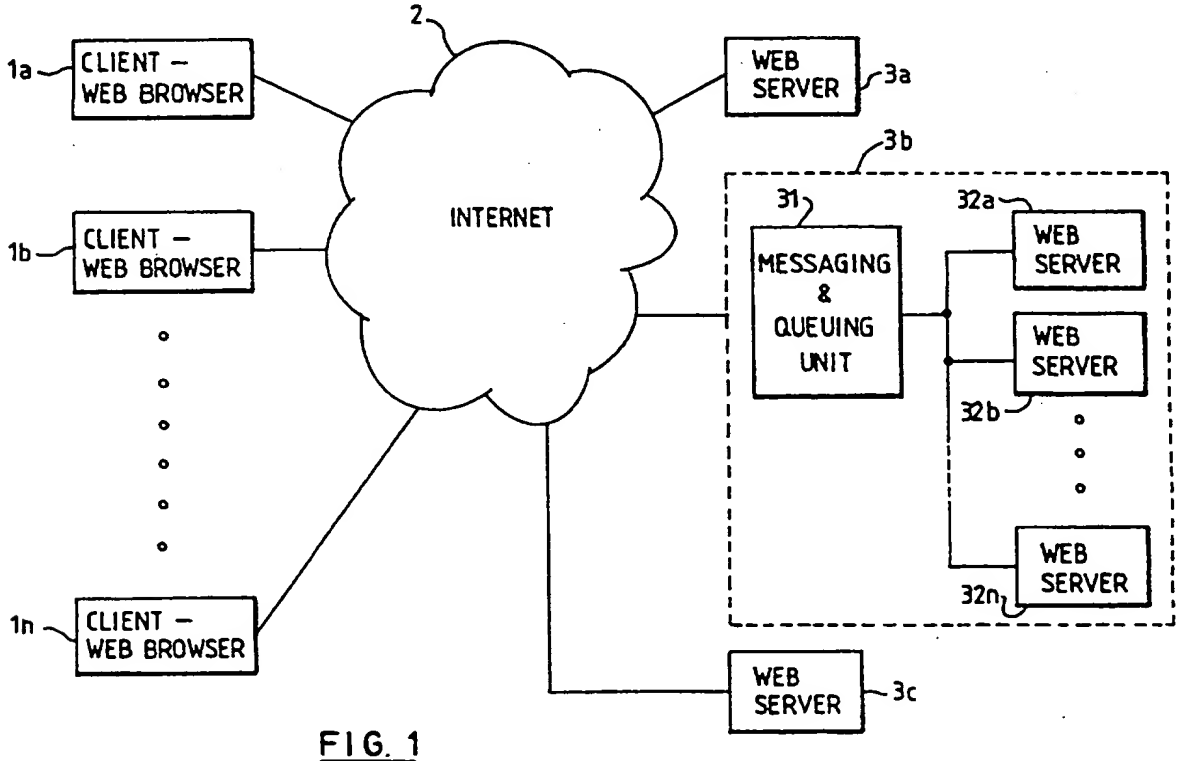

FIG. 1

balancer checks, it may be very busy at a later time in between status checks. In such instances a particular server device can be assigned too much work and thus the respective browsers would have to wait for a long time before receiving the requested information.

Also, with the above known architecture browser requests are taken one at a time by the load balancer and assigned to server devices in the order in which they were received. However, this is disadvantageous because a browser requesting only text would have to wait for a long time while previously received graphics requests are being served (graphics requests involve much more data to be transferred than text because graphics contain much more information than text). Also, if it were particularly more important for one browser user to gain access before the others, there is no mechanism which provided for this in the prior architecture. Each request had to wait its turn.

Also, if there is a particular period of extremely high demand where all available server devices are extremely busy, the browsers are made to wait a long time before having their requests served.

The performance of this architecture is further impaired since each received browser request must be served and a reply sent back to the browser before an initial connection can be made with respect to another browser request.

The present invention has been developed with these limitations in the prior architecture in mind.

Disclosure of the Invention

According to one aspect, the present invention provides a high-availability computer server system capable of serving a large number of requests received from a plurality of computer client devices connected through a network to said server system, said requests specifically identifying said server system, said server system comprising: a messaging and queuing unit having an input connected to said network upon which said requests identifying said server system are received, and an output; and a plurality of server units connected in parallel to said output of said messaging and queuing unit.

By using a messaging and queuing unit, the present invention prevents servers which receive requests from the unit from being overloaded, because the servers "pull" requests off of the queue (in the unit) as opposed to a load balancer "pushing" requests onto the servers without the servers asking for such requests. The server system and thus the overall client/server system thus work much more efficiently to serve client requests, especially in high volume situations where a server system receives a large amount of requests nearly simultaneously.

According to a preferred embodiment, said messaging and queuing unit includes means for assigning priority to received requests. Further, wherein said means for assigning priority assigns higher priority to text requests as compared to graphics requests.

Also, preferably, said messaging and queuing unit includes means for triggering an additional server unit as the number of unserved requests received by said messaging and queuing unit surpasses a threshold amount.

Further, said messaging and queuing means includes means for sending a request to one of said plurality of server units in response to said Web server unit informing said messaging and queuing unit that said Web server unit is ready to serve a request.

Further, the invention preferably provides such a system wherein said network is the World Wide Web, said server system is a Web server system and said client devices are Web browsers.

According to another aspect, the invention provides a method of serving requests received from a plurality of client computer devices via a computer network, each of said requests specifically identifying a specific server system, said method comprising steps of: storing, at said specific server system, said received requests into a messaging and queuing unit; and sending requests from said messaging and queuing unit to a plurality of parallel-connected server units.

Preferably, the method further includes said messaging and queuing unit assigning priority to received requests. Further, wherein said messaging and queuing unit assigns higher priority to text requests as compared to graphics requests.